

Research Article

The Data Big Bang and the Expanding Digital Universe: High-Dimensional, Complex and Massive Data Sets in an Inflationary Epoch

Meyer Z. Pesenson,¹ Isaac Z. Pesenson,² and Bruce McCollum¹

¹IPAC, California Institute of Technology, m/c 220-6, Pasadena, CA 91125, USA

²Department of Mathematics, Temple University, Philadelphia, PA, USA

Correspondence should be addressed to Meyer Z. Pesenson, misha@ipac.caltech.edu

Received 15 June 2009; Revised 21 October 2009; Accepted 12 January 2010

Academic Editor: Joshua S. Bloom

Copyright © 2010 Meyer Z. Pesenson et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recent and forthcoming advances in instrumentation, and giant new surveys, are creating astronomical data sets that are not amenable to the methods of analysis familiar to astronomers. Traditional methods are often inadequate not merely because of the size in bytes of the data sets, but also because of the complexity of modern data sets. Mathematical limitations of familiar algorithms and techniques in dealing with such data sets create a critical need for *new paradigms* for the representation, analysis and scientific visualization (as opposed to illustrative visualization) of heterogeneous, multiresolution data across application domains. Some of the problems presented by the new data sets have been addressed by other disciplines such as applied mathematics, statistics and machine learning and have been utilized by other sciences such as space-based geosciences. Unfortunately, valuable results pertaining to these problems are mostly to be found in publications outside of astronomy. Here we offer brief overviews of a number of concepts, techniques and developments that are vital to the analysis and visualization of complex datasets and images. One of the goals of this paper is to help bridge the gap between applied mathematics and artificial intelligence on the one side and astronomy on the other.

1. Introduction

Astronomy is undergoing a rapid, unprecedented and accelerating growth in both the amount and the intrinsic complexity of data. This results partly from past and future large sky surveys: the Sloan Digital Sky Survey [1], the Large Synoptic Survey Telescope (LSST) [2, 3], ESA's GAIA mission [4], Pan-STARRS [5, 6], the Palomar Transient Factory [7], LAMOST [8], and the Palomar-Quest Survey [9]. Smaller surveys and catalogs comprised of $\sim 10^3$ – 10^4 objects appear annually. The increasing availability of multiple-object spectrographs deployed at ground-based observatories enables observers to obtain spectra of hundreds of objects in a single exposure [10–16]. Million-object spectrographs have been proposed and are undergoing design studies (e.g., [17, 18]). All together, and sometimes individually, such projects are creating for astronomy massive multitemporal

and multispectral data sets comprised of images spanning multiple wavebands and including billions of objects. Furthermore, the Virtual Observatory (VO) is undertaking to combine existing historical data from all wavelengths into what will be, from the user's perspective, a single data set of gigantic size and unprecedented complexity (<http://www.ivoa.net/>).

In addition, mathematically new (for astronomy) forms of data are starting to appear, such as those of the ESA Planck mission in which the cosmic microwave background (CMB) is characterized by a 2×2 matrix at each point in the sky [19].

Other great challenges arise from the so-called three-dimensional (3D) reconstructions. For example, a very important and difficult problem of solar astrophysics is 3D reconstruction of coronal mass ejections [20]. The resulting reconstruction problem cannot be solved via classical methods and must be addressed by more modern

image processing methods like compressed sensing (see Section 5). Another example of tomographic reconstruction is coming from the neutral hydrogen mapping using the redshifted 21 cm line that has recently emerged as a promising cosmological probe. A number of radio telescopes are currently being proposed, planned or constructed to observe the redshifted 21 cm hydrogen line from the Epoch of Reionization (e.g., Fast Fourier Transform Telescope, [21]).

The richness and complexity of new data sets will provide astronomy with a wealth of information and most of the research progress expected from such sets inherently rests in their enormity and complexity. In order to take full advantage of immense multispectral, multitemporal data sets, their analysis should be *automated*, or at least *semi-automated*, and should consist of the following steps: *detection*, *characterization* and *classification* of various features of interest, followed, if necessary, by *automated decision-making* and possibly by *intelligent automatic alerts* (e.g. for follow-up observations or data quality problems). Moreover, a *real-time processing* may be required. Complex data also call for scientific visualization rather than ordinary illustrative visualization. By scientific visualization we mean visualization that does not simply reproduce visible things, but makes the things visible, thus enabling extraction of meaningful patterns from multiparametric data sets and ultimately facilitating analysis. All this requires the development and adaption of modern methods for data representation, analysis and visualization. Methods now standard in astronomy are often ineffective because the data sets are too large and too complex for existing tools to be scaled in a straightforward way from handling several parameters up to dozens or more. There are also important limitations inherent in the mathematical algorithms familiar to astronomers. Scientific visualization, dimensionality reduction, and non-parametric methods in general are among the least-advanced categories of tools in astronomy because until this century there have not been data sets requiring new approaches related to those aspects of the data. All this not only creates a critical need for new sophisticated tools, but moreover urgently calls for *new paradigms* for the analysis, visualization and organization of heterogeneous, multiresolution data across application domains.

The astronomical community is becoming increasingly aware of the fact that new and advanced methods of applied mathematics, statistics and information technology should be developed and utilized. Three of the State of the Profession Position Papers submitted to the Astronomy and Astrophysics Decadal Survey strongly emphasized astronomy's need for new computational and mathematical tools in the coming decade [22–24] (see: http://sites.nationalacademies.org/bpa/BPA_049492). A variety of individual and organizational initiatives and research projects have been proposed or are underway to at least partially address this looming challenge by developing or implementing a new generation of methods of data representation, processing and visualization.

However, there is a large communication gap between astronomy and other fields where adequate solutions exist or are being developed: applied mathematics, statistics and artificial intelligence.

The principal objectives of this paper are twofold. First, we wish to bring attention to some specific needs for new data analysis techniques. Second, we describe some innovative approaches and solutions to some of these problems and give examples of novel tools that permit effective analysis, representation, and visualization of the new multispectral/multitemporal data sets that offer enormous richness if they are mined with the appropriate tools. The extensive amount of relevant work already accomplished in disciplines outside of astronomy does not allow us to offer a complete review of all aspects of these complex topics and problems, but we have selected a number of important examples.

The structure of the paper is as follows. In Section 2 we discuss challenges related to semi-automated processing of low-dimensional images and describe briefly our framework for advanced astronomical image processing. Section 3 describes problems posed by complex astronomical data sets; Section 4 addresses some of these problems and presents some innovative methods for nonlinear dimension reduction, sampling on graphs and manifolds; it also describes briefly a new unifying method for image segmentation and information visualization that is based on physical intuition derived from synchronization of nonlinear oscillations; Section 5 reviews some recent approaches to the challenges posed by high-dimensional, complex data sets and their importance for cosmological and astrophysical problems; in particular, we briefly discuss in this part a new tool called *needlets* for processing data that is laying on the sphere, applications of needlets to the analysis of CMB data, and a generalization of the wavelet-like transforms to Riemannian manifolds. The paper ends with a conclusion.

2. Framework for Processing and Visualizing Astronomical Images

Vast data sets demand automated or semi-automated image processing and quality assessment of the processed images. Indeed, the sheer number of observed objects awaiting analysis makes obvious the need for sophisticated automation of object detection, characterization and classification. Adapting recent advances of computer vision and image processing for astronomy and designing and implementing an image processing framework (see below) that would utilize these continuing achievements, remains, however, a major challenge.

Developers and users alike have realized that there is more to creating an application than simple programming. The objective of creating a flexible (see below) application is customarily achieved by exploiting the object-oriented paradigm [25, 26]. In computer science such systems are called frameworks [27]. Frameworks are flexible, and users can extend frameworks' capabilities by installing plug-ins without having to rewrite the basic code. An application framework consists of computational (processing) modules, data, and an interactive interface. Frameworks are extendable (i.e., they can easily adopt software products to changes of specification) and reusable (the software elements are able to serve for construction of many different new applications).

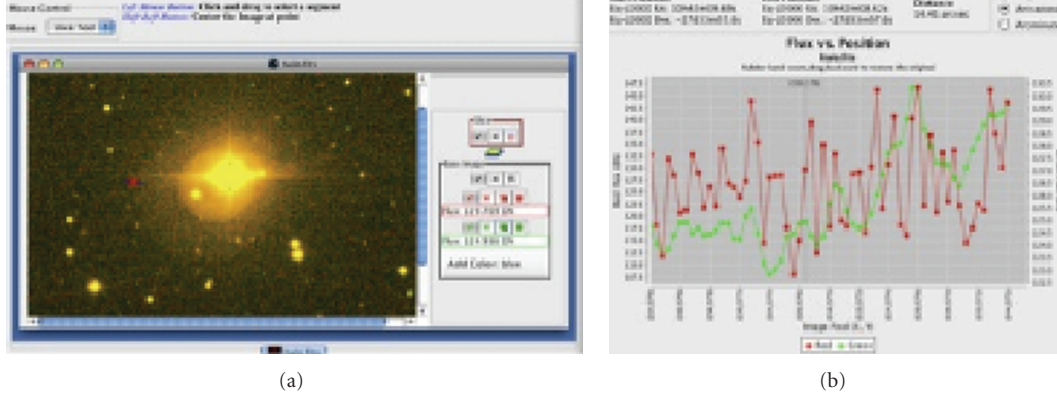


FIGURE 1: ESO-La Silla; courtesy of A. Grado, INAF-Osservatorio Astronomico di Capodimonte. (a) overlaid pre- and post-processed images; the red cross shows approximately the edge of the diffuse halo. (b) Flux-cut through the overlaid pre- and post-processed images (red and green, resp.); the vertical grey line (close to the center of the plot) corresponds to the red cross in the plot on the left; after the preprocessing the average level “outside” of the halo is lower than the average level “inside”, thus enabling a better automated separation of the diffuse halo from the background.

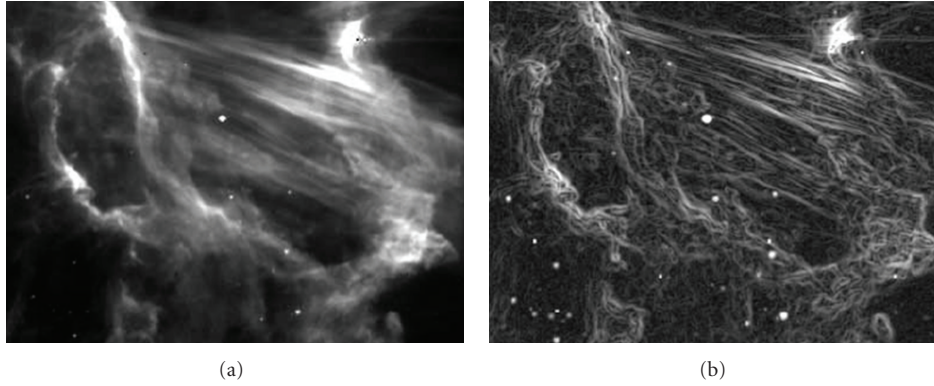


FIGURE 2: Morphology unveiling [28]. (a) IC 405, *Spitzer* IRAC camera $8.0\mu\text{m}$ image (France et al. [29]); filaments and a bow shock near HD34078 (see the red cross “x” in Figure 3(a)). The brightness is proportional to the flux (logarithmic scale). (b) Brightness is proportional to the module of the gradient of flux (logarithmic scale); this way of looking at astronomical images facilitates analysis of nebular morphology, outflows, jets, embedded sources, and shock fronts.

The keystone elements of a system that unifies a wide range of methods for astronomical image processing should be computational and visualization modules. Such a platform-independent framework with an integrated environment was described in Pesenson et al. [28]. It provides many common visualization features plus the ability to combine these images (overlays, etc.) in new and unique ways. The framework has a very intuitive GUI that allows the user to set up a custom pipeline interactively and to process images singly or in a batch mode. The final products, as well as the results of each step, are viewable with the framework. It also provides access to different data archives and can easily incorporate custom modules written in any programming language. Figures 1–4 give a few self-explanatory examples that demonstrate some of the framework’s functionality.

The framework deals primarily with image regularization and segmentation. These are fundamental first steps for the detection and characterization of image elements or objects and as such they play principal roles in the realization

of automated computer vision applications. Image segmentation can roughly be described as the process that subdivides an image into its constituent parts (objects) and extracts those parts of interest. Since the inception of image segmentation in the 1960s, a large number of techniques and algorithms for image segmentation have been developed.

However, due to revolutionary advances in instrumentation, the complexity of images has changed significantly: the extension of grey level images to multi- and hyperspectral images, from 2D images to 3D, from still images to sequences of images, tensor-valued images (polarization data), and so forth. Some modern, cutting-edge methods for image processing have been developed lately, and are being developed today, by information scientists outside of astronomy. The substantial progress in this direction made by the image processing and computer vision communities [30, 31], has found multiple applications in physics, technology, and biomedical sciences. Unfortunately, for the most part these advances have not yet been utilized by the astronomical community.

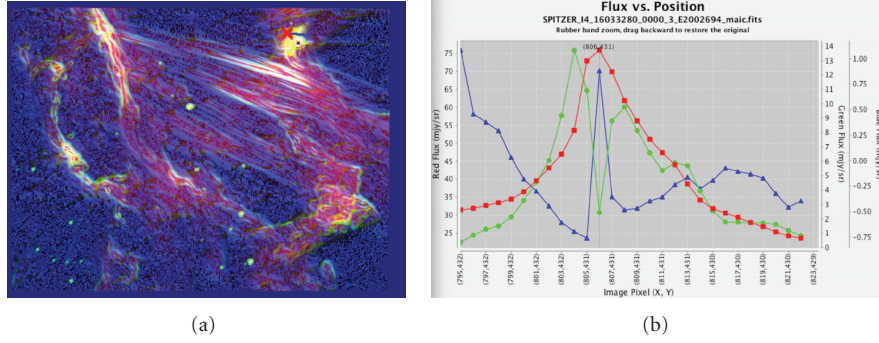


FIGURE 3: Overlaid pre- and postprocessed images [28]. (a) Three overlaid images of IC 405: *Spitzer* IRAC camera 8.0 μm image, the module of its gradient, and the angle of its gradient (red, green, and blue, resp.). The line through the red cross “x” indicates the crosscut. (b) Three profiles for the crosscut through “x”. The crosshair goes through the ridge point and a local minimum of the module of the gradient. Note the different scales on the left and right vertical axes.

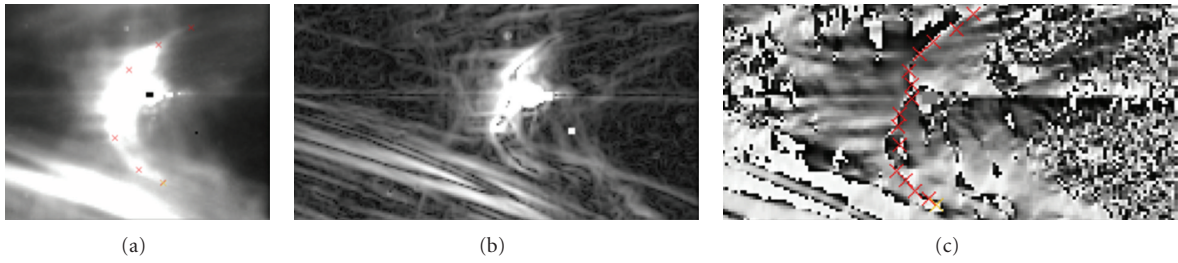


FIGURE 4: Three different visualizations of the bow shock structure [28]. (a) IC 405, *Spitzer* IRAC camera 8.0 μm image, the bow shock (red crosses) near HD 34078. The brightness is proportional to the flux (logarithmic scale). (b) The bow shock area where the brightness is proportional to the module of the gradient of flux (logarithmic scale). (c) The bow shock area where the front of the bow shock is immediately apparent as a curve of 1 pixel width (red crosses); the brightness is proportional to the angle of the gradient of flux (logarithmic scale).

Multiscale image representation and enhancement are such approaches. They have become important parts of computer vision systems and modern image processing. The multiscale approach has proven to be especially useful for image segmentation for feature, and artifact detection. It enables a reliable search for objects of widely different morphologies, such as faint point sources, diffuse supernova remnants, clusters of galaxies, undesired data artifacts, as well as unusual objects needing detailed inspection by a scientist. It is well known that in astronomical images one often sees both point sources and extended objects such as galaxies embedded in extended emission (see, e.g., Figures 2, 6 and 7). Because of the issue of robustness with respect to noise (Section 4), a careful preprocessing is required before one can safely apply image segmentation or dimension reduction. So, an adequate way of preprocessing is what needs to be addressed first.

One effective approach to denoising is based on partial differential equations and may be seen as the local adaptive smoothing of an image along defined directions that depend on local intensities. One wants to smooth an image while preserving its features by performing a local smoothing mostly along directions of the edges while avoiding smoothing orthogonally to these edges. Many regularization schemes have been developed so far for the case of

simple two-dimensional scalar images. An extension of these algorithms to vector-valued images is not straightforward. For a gray-scale image, the gradient is always perpendicular to the level set objects of the image; however, in the multi-channel case, this quality does not hold. Applying nonlinear diffusion to each channel or spectral band separately is one possible way of processing multi- and hyperspectral cubes; however, it does not take advantage of the richness of the multi/hyperspectral data. Moreover, if the edge detector acts only on one channel, it may lead to an undesirable effect, such as color blurring, where edges in color images are blurred due to the different local geometry in each channel. Hence, a coupling between image channels should appear in the equations through the local vector geometry. We achieve this by implementing a nonlinear diffusion on a weighted graph, thus generalizing the approach adopted by Pesenson et al. [28]. The governing equation is linear, but the nonlinearity enters through the weights assigned to the graph edges. This algorithm respects the local smoothing geometry and thus serves well as a preprocessing step required for dimension reduction (Section 4).

Our framework handles two-dimensional scalar images and paves the way to the semi- and automated image processing and image quality assessment. However, the ability to extract useful knowledge from high-dimensional

data sets is becoming more and more important (Section 3). It is closely related to finding complicated structural patterns of interrelated connections and heavily depends on the ability to reduce dimensions of the raw data. This problem constitutes a big challenge for the scientific and technology communities. To extend the framework's functionality to high-dimensional images and data sets we have been developing novel, practical algorithms for dimension reduction. In the next section we describe some of the challenges presented by modern, high-dimensional data sets.

3. Complex, Massive Data Sets

Astronomy has long found the use of multiple data dimensions to be crucial aids to progress. For example, surveys of H-alpha sources when cross-correlated with spectral types guided astronomy to discover interesting new types of objects such as Herbig AeBe stars [32]. Cross-comparison of brightnesses, redshifts, and optical object morphologies led to the discovery of quasars. X-ray emission proved to be a highly efficient method of identifying pre-main sequence objects in a large field of view (e.g., review [33]). The fact that the same object (SS 433) was noticed to exhibit H-alpha emission and strong radio and optical variability led astronomers to investigate further and thus discover the existence of microquasars [34–36]. Other discoveries from the use of multidimensional data using just a few dimensions include such things as Be X-ray binaries and soft gamma ray repeaters.

Simple color-magnitude diagrams are another traditional tool taking advantage of multiple data dimensions (e.g., cataloging YSO candidates from Spitzer survey data, Whitney et al. [37]). In addition, it is well known that simple color-color plots using four colors provide ways to efficiently and fairly reliably classify into physical types a large number of objects (IRAS data, e.g., [38, 39]; 2MASS data, e.g., [40] MSX data, e.g., [41, 42]; *Spitzer* data, e.g., [43]) as well as to discover interesting new objects as outliers (e.g., Luminous Red Novae, [44, 45]). The meaning and usefulness of more complicated multiwavelength cross-correlations across widely-separated wavelength domains remain a nontrivial but fruitful challenge still being explored (e.g., microquasars identified by comparing radio, IR, and X-ray properties [46]).

As multiwavelength data have become available for huge numbers of objects in the past few decades, the number of data dimensions for a typical object has grown beyond what can be visualized and studied using classical color-color plots and correlations using only a few dimensions. For example, Egan et al. [41] write “six-dimensional data are difficult to represent; in addition, it is not clear that all of the colors yield completely independent information” (see Figure 5).

While Figure 5 successfully permits one to identify meaningful patterns, it is easily imagined that attempting to display a much larger number of data dimensions by simply using many more colors and symbols overlaid on the same plot and same region of parameter space would not be a useful form of scientific visualization. Also, two-dimensional

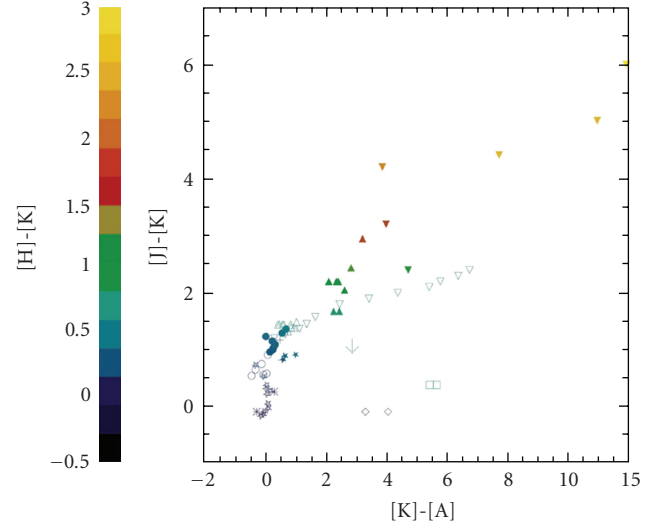


FIGURE 5: Color-color plots are a familiar form of scientific visualization. This example (from [41]) shows, simultaneously, infrared color-color information for several different categories of objects. Different object classes are represented by different symbols or different hues. Thus several data dimensions (the four IR colors and the several classes of object) are represented.

plots are obviously not useful for expressing relationships among more than four colors. Effective scientific visualization of a large number of data dimensions requires new techniques.

The catalog information available for a great many objects already includes such information as magnitude, several optical and IR colors, metallicities, spectral types, and so on. A galaxy catalog would of course include other parameters such as morphology and redshift. When spectra are considered and compared in detail, the huge number of emission and absorption features obviously compounds the problem vastly. Still more complexity is added when one attempts to correlate a large grid of models with a large data set having many dimensions (e.g., the YSO analysis [37]) in order to create feedback for models based on statistically significant samples rather than on a few putative prototypes.

Other examples of high-dimensional data include, but are not limited to, multiparametric data sets (e.g., the manifold of galaxies, [47, 48]), multitemporal, multispectral and hyperspectral data sets and images, and high-resolution simulations on massively parallel computers. The magnitude of the computational challenge for pattern recognition and classification algorithms is suggested by the fact that the VO will contain billions of data points each having hundreds of dimensions [49].

All these examples clearly demonstrate that the automated and semiautomated processing required by the unprecedented and accelerating growth in both the amount and the complexity of astronomical data demands new ways of information representation. In the next section we discuss such approaches and describe some of the original algorithms we have developed in the course of this ongoing work.

4. Dimensionality Reduction and Sampling for Complex and Large Data Sets

Because approaches to complex data require advanced mathematics, astronomers who wish to take advantage of them will need at least some basic knowledge of new, unfamiliar mathematical concepts and terminology. The full practical adoption of such methods requires interdisciplinary scientists who understand the new approaches in depth and are interested in working with astronomers to adapt and apply the methods to astronomical data sets. This is already happening as part of some research. The first step for a “neophyte” astronomer, however, is to learn “what is out there” as a basis for further investigation and consideration of the utility of various methods. The goal of this section is to offer a very basic introduction and explanation of some of these unfamiliar concepts.

Machine learning [50] is becoming increasingly important in astronomy (see extensive reviews [51, 52]). The main objectives of machine learning are clustering (automatic identification of groups of similar objects), and classification (assigning labels to instances). However, high dimensionality complicates machine learning and can easily thwart the entire effort. It also becomes a formidable obstacle in computing numerical solutions in Bayesian statistics for models with more than a few parameters.

Bellman [53] coined the term “curse of dimensionality”, to describe how difficult it was to perform high-dimensional numerical integration. For example, 100 evenly-spaced sample points suffice to sample a unit interval with no more than 0.01 distance between points; an equivalent sampling of a 10-dimensional unit hypercube with a lattice with a spacing of 0.01 between adjacent points would require 10^{20} sample points. Thus, in some sense, the 10-dimensional hypercube can be said to be a factor of 10^{18} “larger” than the unit interval. *Obviously, this will make many computational tasks intractable for high-dimensional data sets.*

Euclidian spaces are usually used as models for traditional astronomical data types (scalars, arrays of scalars). Another *general manifestation* of high dimensionality is the fact that in a high-dimensional Euclidean space, volume expands far more rapidly with increasing diameter than it expands in lower-dimensional spaces. Indeed, if one compares the size of the unit ball with the unit cube as the dimension of the space increases, it turns out that the unit ball becomes an insignificant volume relative to that of the unit cube. Thus, in some sense, nearly all of the high-dimensional space is “far away” from the center. This is called the “empty space phenomenon” [54, 55]: the high-dimensional unit space can be said to consist almost entirely of the “corners” of the hypercube, with almost no “middle”. One more important example of the unexpected properties of high-dimensional Euclidean spaces is the following behavior of the Gaussian distribution in high dimensions: the radius of a hypersphere that contains 95% of the distribution grows as the dimensionality increases.

These sorts of problems demonstrate that in order to make practical the extraction of meaningful structures

from multiparametric, high-dimensional data sets, a low-dimensional representation of data points is required. Dimension reduction (DR) is motivated by the fact that the more we are able to reduce the dimensionality of a data set, the more regularities (correlations) we have found in it and therefore, the more we have learned from the data. Data dimension reduction is an active branch of applied mathematics and statistics [50]. It consists of methods for finding lower-dimensional representation of high-dimensional data, without losing a significant amount of information, by constructing a set of basis functions that capture patterns intrinsic to a particular state space. DR methods greatly increase computational efficiency of machine learning algorithms, improve statistical inference and enables effective scientific visualization and classification. From a large set of images obtained at multiple wavebands, effective dimension reduction provides a comprehensible, information-rich single image with minimal information loss and statistical details, unlike a simple coadding with arbitrary, empirical weights (see a simple four-wavelength example in Figures 6 and 7).

Classical approaches to dimension reduction not unfamiliar to astronomy are principal components analysis (PCA) [56] and multidimensional scaling [57]. Although first applied to astronomy by Bijaoui [56] in 1974, PCA was not commonly used until the 1990’s [58, 59]; see also a general introduction to PCA in [60]. It has been used for such things as classification of galaxies and quasars [59, 61], photometric and spectroscopic redshift estimation (e.g., [62, 63], sky subtraction [64] and optical spectral indicators [65]. PCA has been used for the simultaneous analysis of dozens of data parameters for each member of a sample of 44 active galactic nuclei [66], which appears to be the largest number of parameters yet analyzed using PCA in an astronomical study.

PCA has a serious drawback in that it does not explicitly consider the structure of the manifold on which the data may possibly reside. In differential geometry an n -dimensional manifold is a metric space that on a small scale resembles the n -dimensional Euclidean space; thus a circle is a one-dimensional manifold, while a sphere is a two-dimensional manifold. PCA is intrinsically linear, so if data points form a nonlinear manifold, then obviously, there is no rotation & shift of the axis (this is what a linear transform like PCA provides) that can “unfold” such manifold. In other words, if the data are mainly confined to an almost linear low-dimensional subspace, then simple linear methods such as PCA can be used to discover the subspace and estimate its dimensionality. If, on the other hand, the data lie on (or near) a highly nonlinear low-dimensional submanifold, then linear methods are not effective in capturing the fine meaningful structures in the data. A blind application of linear methods may result in a complete misrepresentation of the data.

We have recently developed some advanced, original methods for performing nonlinear DR, which do not suffer from the limitations of PCA. In what follows, we briefly describe these methods. First, we introduce some more concepts and methods that have proved to be effective in the area of machine learning.

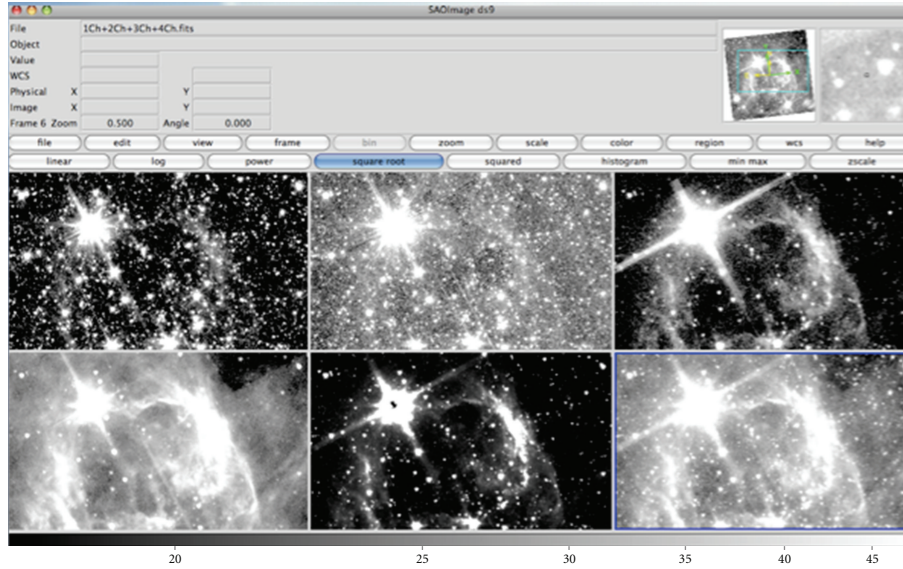


FIGURE 6: *Spitzer* IRAC camera $8.0\ \mu\text{m}$ image, 1–4 channels. Upper row: channels 1–3. Lower row: channel 4; the result of applying dimension reduction; sum of the channels 1–4.

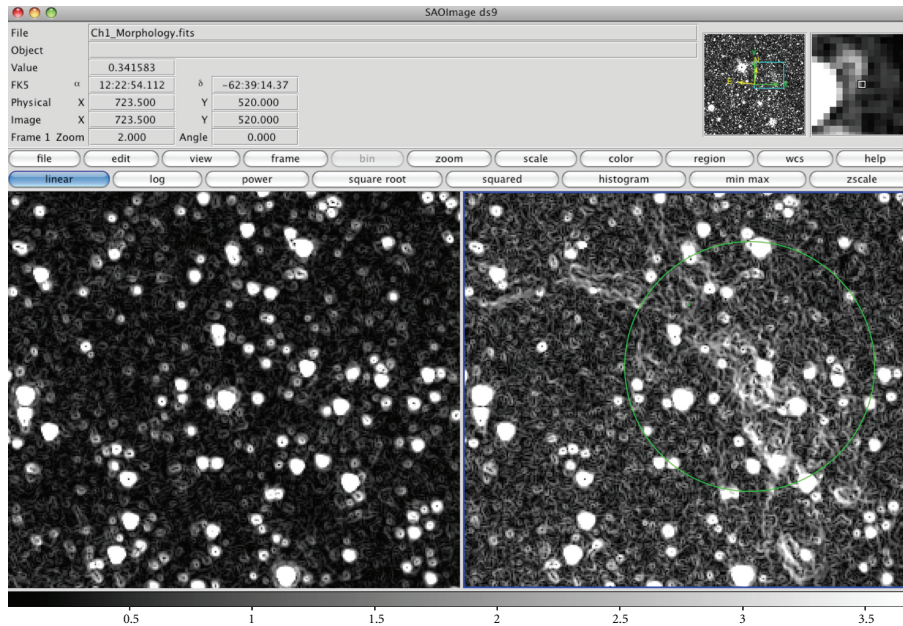


FIGURE 7: *Spitzer* IRAC camera $8.0\ \mu\text{m}$ image, Left: channel 1, morphology unveiling following [28]. Right: morphology unveiling performed on the reduced image from Figure 6. It demonstrates that dimensionality reduction retains the important features from all channels. This is especially useful for dealing with large numbers of images.

4.1. Graphs. In the context of data retrieval and processing, dimensionality reduction methods based on *graph theory* have proved to be very powerful. In mathematics and computer science, graph theory is the study of mathematical structures to model relations between objects [67]. Graph theory has been successfully applied to a wide range of very different disciplines, from biology to social science, computing and physics. Among many applications, graphs and manifolds have been used to address mining databases, Internet search engines, computer graphics, computation reordering, image processing, and so forth. In what follows

we discuss various data structure and algorithms that we have already developed and are developing.

The graph representation of structured data provides a fruitful model for the relational data mining process. A graph is a collection of nodes and links between them; the nodes represent data points and the weights of the links or edges indicate the strength of relationships. A graph in which each graph edge is replaced by a directed graph edge is called a directed graph, or diagraph [67]. Diagraphs are used for context-sensitive browsing engines and for ranking hyper-linked documents for a given query.

The modern approach to multidimensional images or data sets is to approximate them by graphs or Riemannian manifolds. The first important, and very challenging, step is to convert such a data cloud to a weighted finite graph. The next important problem is the choice of “right” weights that should be assigned to edges of the constructed graph. The weight function describes a notion of “similarity” between the data points and as such strongly affects the analysis of the data. The weights should entirely be determined by application domain. The most obvious way to assign weights is to use a positive kernel like an exponential function whose exponent depends on the local Euclidean distance between data points and a subjectively chosen parameter called “bandwidth”. (There are also other ways of assigning weights, which depend on more complex mathematics than we discuss in this article.)

Next, after constructing a weighted graph, one can introduce the corresponding combinatorial Laplace operator [67, 68]. In machine learning, methods based on graph Laplacians have proved to be very powerful [50]. The eigenfunctions and eigenvalues of the Laplacian form a basis, thus allowing one to develop a harmonic or Fourier analysis on graphs. By further developing harmonic analysis on graphs and manifolds [69–81], we have devised innovative algorithms for data compression and nonlinear data dimension reduction. These results enable one to overcome PCA’s limitations for handling nonlinear data manifolds and also allow one to deal effectively with incomplete data (such as missing observations or partial sky coverage).

4.2. Hypergraphs. Most existing data mining and network analysis methods are limited to pairwise interactions. However, sets of astronomical objects usually exhibit multiple relationships, so restricting analysis to the dyadic (pairwise) relations leads to loss of important information and to missing discoveries. Triadic, tetradic or higher interactions offer great practical potential. This has led to the approach based on hypergraphs (e.g., [82]). Hypergraphs constitute an important extension of graphs that allow edges to connect more than two vertices simultaneously. Thus hypergraphs provide a more comprehensive description of feature relations and structures. Hypergraphs furnish a much more adequate approach to real world data sets and allow one to deal with clustering and classifications using higher order relations. It has been shown that, in general, there does not exist a graph model that correctly represents the cut properties of the corresponding hypergraph (Ihler et al. [83]). Thus new mathematical methods are required to take advantage of the richness of information available in hypergraphs. In order to provide means for analysis of databases with multiple relationships, we are currently developing extensions of the original methods described in the previous subsection.

4.3. Fractals. Data sets having fractional dimensions (“fractals”: Mäkelä [84]; Schroeder [85]; Faloutsos [86]) have been suggested to represent many phenomena in astronomy.

Examples include star formation on galactic scales [87], the stellar initial mass function [88], galaxy distributions in space [89, 90], cloud distributions in active galactic nuclei [91], the cosmic microwave background [92], thermonuclear flame velocities in Type Ia supernovae [93], and the mass spectrum of interstellar clouds [94]. Such data sets exhibit a dimensionality which is often much lower than the dimension of the Euclidean space they are embedded into. For example, the distribution of galaxies in the universe has dimension $D \sim 1.23$. The difference between the two dimensions occurs because the fractal dimensionality is intrinsic. The intrinsic dimension of a graph reflects the intrinsic dimension of the sampled manifold.

Obviously, an important first step in practical dimensionality reduction is a good estimate of the intrinsic dimension of the data. Otherwise, DR is no more than a risky guess since one does not know to what extent the dimensionality can be reduced. To enable analysis of astronomical data sets that exhibit a fractal nature, we are currently developing a practical concept of spectral dimensionality, as well as original algorithms for sampling, compression and embedding fractal data.

4.4. The Petascale Connection. The approaches described above dealt with compact manifolds and finite graphs. However, massive data sets are more adequately described by noncompact manifolds and infinite graphs. In order to deal with extremely large data sets we extended dimension reduction to *non-compact* manifolds and *infinite* graphs. We are also working on generalizing the Fourier analysis to *non-compact* Riemannian manifolds and *infinite* quantum and combinatorial graphs, directed graphs, hypergraphs and some fractals. Implementation of these algorithms, and incorporation of them into the framework described in Section 2, will enable a more adequate analysis of massive data sets.

4.5. Robustness of Dimension Reduction Algorithms with Respect to Noise. Despite the important and appealing properties of the above mentioned dimension reduction algorithms, both linear and nonlinear approaches are sensitive to noise, outliers, and missing measurements (bad pixels, missing data values). Because noise and data imperfections may change the local structure of a manifold, locality preservation means that existing algorithms are topologically unstable and not robust against noise and outliers [95].

This is obviously a serious drawback because astronomical data are always corrupted by noise. Budavári et al. [96] presented a robust version of PCA and applied it to astronomical spectra. Their approach addressed the issues of outliers, missing information, large number of dimensions and the vast amount of data by combining elements of robust statistics and recursive algorithms that provide improved eigensystem estimates step by step. However, as it was mentioned earlier, the PCA method is intrinsically linear and cannot be applied to data points residing on a nonlinear manifold.

Thus the practical usage of dimension reduction demands careful improvement of signal-to-noise ratio without smearing essential features. Implementing a nonlinear diffusion on weighted graphs (Section 2) enables us to apply dimension reduction to noisy data sets as well.

4.6. Image Segmentation, Unsupervised Learning and Information Visualization. In what follows we briefly describe a new, original unifying approach to segmentation of images in particular and pattern recognition and information visualization in general. Image segmentation (see Section 2) plays a principal role in the realization of automated computer processing, as a previous stage for the recognition of different image elements or objects.

Solutions to the manifold learning problem can be based on intuition derived from physics. A good example of this is the approach of Horn and Gottlieb [97] that is based on the Schrödinger operator, where they constructed an approximate eigenfunction, and computed its corresponding potential. The clusters were then defined by the minima of the potential.

We developed [98] an alternative that is also based on physical intuition, this one being derived from synchronization of nonlinear oscillations [99, 100]. Approximating a multidimensional image or a data set by a graph and associating a nonlinear dynamical system with each node enables us to unify the three seemingly unrelated tasks: image segmentation, unsupervised learning and data visualization. Pattern recognition may benefit significantly from new methods of visualization and representation of data, since they may uncover important relationships in multidimensional data sets. Bringing out patterns by setting data points into oscillatory motion is a very effective way of visualizing data [101]. Our method is a feature-based, multilevel algorithm that finds patterns by employing a nonlinear oscillatory motion. At the same time, the oscillatory motion reveals to the eye patterns by making them oscillate coherently with a frequency different from the rest of the graph. Patterns are detected recursively inside the graph, and the found features are either collapsed into single nodes, forming a hierarchy of patterns or can be zoomed in and studied individually. This method can be described as a precomputed animation and it enables both qualitative (by eye) and quantitative discovery of correlations (see Figure 8).

The approaches presented in this section enable interpolation, smoothing and immersions of various complex (dozens or hundreds of useful parameters associated with each astronomical object) and large data sets into lower-dimension Euclidean spaces. Classification in lower dimensional space can be done more reliably than in high dimensions. Thus, DR can be significantly beneficial as a pre-processing step for many existing astronomical packages, such as for example, the popular source extractor SExtractor [102]. Incorporating into our framework (Section 2) the graph/manifold-based tools described in this section, will allow to address multiple pattern recognition (clustering) tasks as well as visualization of multidimensional data. Altogether, these approaches provide important generalizations

of the tools for spectral clustering and dimensionality reduction, and enable more adequate representation, effective data retrieval and analysis of complex, modern astronomical data sets.

5. Some Recent Approaches to the Challenges of Data Intensive Astronomy

Data intensive astrophysics requires an interdisciplinary approach that will include elements of applied mathematics [67, 103], modern statistical methods [104, 105], machine learning [50], computer vision [31] and image processing [30]. The breadth and complexity of the work relevant to modern astronomical data challenges is large, and does not permit a full treatment here. However, it is valuable to briefly mention a number of important problems, approaches and efforts that have been pursued.

The problem of processing data that lay on a manifold is very important for cosmological data analysis. The standard, powerful data analysis package HealPix [106] processes data on a two-dimensional manifold—the sphere. The concept of needlets (second generation spherical wavelets) has recently attracted a lot of attention in the cosmological literature. The first application of needlets to cosmological data was provided by Pietrobon et al. [107]. They analyzed cross-correlation of Wilkinson Microwave Anisotropy Probe (WMAP) CMB data with NRAO VLA Sky Survey (NVSS) radio galaxy data. The approach based on needlets enabled more accurate statistical results related to the dark energy component. The investigation of needlets from the probabilistic viewpoint and their relevance for the statistical analysis of random fields on the sphere was done for the first time by Baldi et al. [108]. A thorough presentation of the spherical needlets for CMB data analysis is given by Marinucci et al. [109]. Various issues related to CMB, such as spectrum estimation, detection of features and anisotropies, mapmaking were addressed by Faÿ et al. [110], Pietrobon et al. [111] (see also references there). The CMB models are best analyzed in the frequency domain, where the behavior at different multipoles can be investigated separately; on the other hand, such problems as missing observations or partial sky coverage make impossible the evaluation of spherical harmonic transforms. The needlets allow for a very efficient simple reconstruction formula that enables one to perform frequency analysis by using only partial information about data and providing means for handling masked data. Besides, needlets exemplify other important properties that are not generally shared by other spherical wavelet constructions: they do not rely on any kind of tangent plane approximation; they have good localization properties in both pixel and harmonic space; and needlet coefficients are asymptotically uncorrelated at any fixed angular distance (which makes their use in statistical procedures very promising). All these unique properties make needlets a very valuable tool in various areas of CMB data analysis. Recently, Geller and Marinucci [112] introduced spin needlets as a tool for the analysis of spin random fields. Geller et al. [113] adopted the spin needlet approach for the analysis of CMB polarization

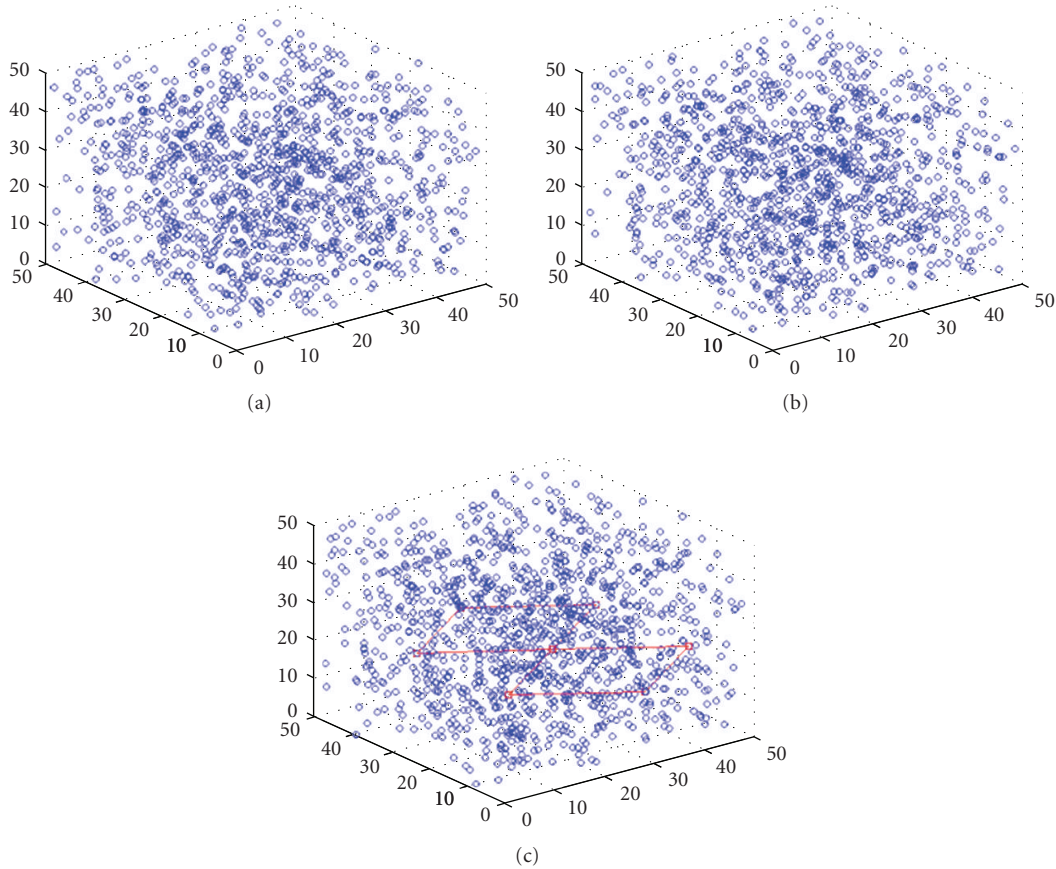


FIGURE 8: Testing the algorithm on synthetic data. A simulated three-dimensional set of a thousand uniformly distributed random points with a double-diamond pattern created by assigning large weights to the edges connecting the points in the pattern (while the rest of the weights are negligible). (a) and (b): two screen shots from a running animation—each point in the set oscillates (in this case in three dimensions) with its own, random frequency. (c) Synchronization of the points that are connected with high-weight edges allows to reveal the pattern visually (to avoid extra clutter, the edges are not displayed in the animation), or automatically, by selecting synchronized points and highlighting them.

measurements. Lately Geller and Mayeli [114] constructed continuous wavelets and nearly tight frames (needlets) on compact manifolds.

The essential part of the general analysis based on the wavelet-like constructions is sampling of bandlimited functions. The mathematical foundations of sampling on arbitrary Riemannian manifolds of bounded geometry were laid down by Pesenson [115, 116]. Recently Geller and Pesenson [117], constructed bandlimited and highly concentrated tight frames on compact homogeneous manifolds. These results can be considered as an extension of the wavelet-like transforms to Riemannian manifolds, thus enabling one to reconstruct information from incomplete data defined on an arbitrary manifold.

Multiscale data analysis has proved to be a very powerful tool in many fields. Applications of multiscale image analysis to astronomy were pioneered by Starck et al. [118] and Starck and Murtagh [119].

Spectral methods and diffusion maps have recently emerged as effective approaches to nonlinear dimensionality reduction [120–122]. The diffusion maps approach has

successfully been applied to analysis of astronomical spectra by Richards et al. [123].

Manifold learning may be seen as a DR procedure aiming at capturing the degrees of freedom and structures (clusters, patterns) within high-dimensional data. Manifold learning nonlinear algorithms such as isometric mapping (ISOMAP) by Tenenbaum et al. [124] and local linear embedding (LLE) by Roweis and Saul [125] project high-dimensional manifold data into a low-dimensional space by preserving the local geometric features. Ball and Brunner [51] provide a very broad review of the current state of machine learning and data mining in astronomy.

The AstroNeural collaboration group [52] implemented tools based on neural networks, fuzzy-C sets and genetic algorithms, and applied them to perform complex tasks such as unsupervised and supervised clustering and time series analysis. D’Abrusco et al. [126] present a supervised neural network approach to the determination of photometric redshifts. D’Abrusco et al. [127] describe a method for the photometric selection of candidate quasars in multiband surveys based on the probabilistic principal surfaces. Pasian

TABLE 1: Examples of complex data types and some of the methods for their representation and processing.

Data Types	Some Astronomical Applications	Traditional Approaches to Data Representation & Processing	Advanced Approaches to Data Representation & Processing
Vector Data	(1) Multiwavelength observations.	(1) Linear dimension reduction: PCA and its modifications.	(1) Spectral methods, eigenmaps, diffusion maps, LLE, ISOMAP.
	(2) Multitemporal observations.		(2) Sampling on graphs.
	(3) VO.		(3) Methods based on nonlinear dynamics.
	(4) Spectra.		(4) Neural networks, fuzzy-C sets. (5) Genetic algorithms. (6) Scientific visualization. (7) Compressed sensing.
Manifold-Valued and/or Manifold-Defined	(1) Polarization measurements (CMB).	(1) Various sampling distributions on the sphere.	(1) Healpix (data on 2D sphere).
	(2) Gravitational lensing.		(2) Needlets.
	(3) Solar astrophysics.		(3) Sampling on manifolds.
			(4) Scientific visualization.

et al. [128] give a general review of the development of tools to be subsequently used within the international VO.

Comparato et al. [129] show how advanced visualization tools can help the researcher in investigating and extracting information from data. Their focus is on VisIVO, a new open-source graphics application that blends high performance multidimensional visualization techniques and up-to-date technologies to cooperate with other applications and to access remote, distributed data archives.

Draper et al. [130] discuss the GAIA application for analyzing astronomical image and show how the PLASTIC protocol has been used to inter-operate with VO enabled applications.

The Center for Astrostatistics (CASt) at Pennsylvania State University provides a wealth of resources (codes, data, tutorials, programs, etc.) related to challenges in statistical treatments of astrophysical data: <http://astrostatistics.psu.edu/>.

A large amount of practical and up-to-date information (texts, tutorials, preprints, software, etc.) related to Bayesian inference in astronomy and other fields is provided by T. Loredo on the website Bayesian Inference for the Physical Sciences (BIPS) at <http://www.astro.cornell.edu/staff/loredo/bayes/index.html>.

The International Computational Astrostatistics (InCA) Group at Carnegie Mellon University develops and applies new statistical methods to inference problems in astronomy and cosmology, with an emphasis on computational non-parametric approaches (see for details <http://www.stat.cmu.edu/~inca/index.html>).

The AstroMed project at Harvard University's IIC is dedicated to the application of medical image visualization to 3D astronomical data [131].

Compressed sensing and the use of sparse representations offer another promising new approach. Traditionally it has been considered unavoidable that any signal must be sampled at a rate of at least twice its highest frequency in

order to be represented without errors. However, a technique called compressed sensing that permits error-free sampling at a lower rate has been the subject of much recent research. It has great promise for new ways to compress imaging without significant loss of information, thus ameliorating analysis limitations deriving from limitations of computing resources. The power of compressed sensing was strikingly illustrated when an object was successfully imaged in some detail by a camera composed of a *single pixel* [132]. Bobin et al. [133] discuss recent advances in signal processing that use sparse representations. Relatively nontechnical introductions to compressed sensing may be found in Candes and Wakin [134] and Romberg [135]. The first astronomical results in the literature based on the use of compressed sensing are that of Wiaux et al. [136].

Various data types together with methods used for their representation are briefly summarized in Table 1. The table is not comprehensive, but it provides a quick overview of what was discussed above.

6. Conclusion

Extremely large data sets, as well as the analysis of hundreds of objects each having a large number of data dimensions, present astronomy with unprecedented challenges. The challenges are not only about database sizes in themselves, but about how intelligently one organizes, analyzes, and navigates through the databases, and about the limitations of existing data analysis approaches familiar to astronomy. The answers to these challenges are not trivial, and for the most part lie in complex fields of research well outside the training and expertise of almost all astronomers. Fortunately, other disciplines such as imaging science and earth sciences have for many years been grappling with the same sorts of problems. Fruitful interdisciplinary work has already become a regular feature of research in those other disciplines, and has resulted in applications of crucial value to other sciences

seeking to take advantage of complex, giant data sets in their respective fields. This work has brought about many helpful applications and promising paths for further progress that potentially have significant value to astronomy.

Multidimensional image processing, image fusion (combining information from multiple sensors in order to create a composite enhanced image) and dimension reduction (finding lower-dimensional representation of high-dimensional data) are effective approaches to tasks that are crucial to multi-temporal, multiwavelength astronomy: study of transients, large-scale digital sky surveys, archival research, and so forth. These methods greatly increase computational efficiency of machine learning algorithms and improve statistical inference, thus facilitating automated feature selection, data segmentation, classification and effective scientific visualization (as opposed to illustrative visualization). Dimensionally reduced images also offer an enormous savings in storage space and database-transmission bandwidth for the user, without significant loss of information, if appropriate methods are used.

To effectively use the large, complex data sets being created in 21st Century astronomy, significant interdisciplinary communication and collaboration between astronomers and experts in the disciplines of applied mathematics, statistics, computer science and artificial intelligence will be essential. The concepts and approaches described in this paper are among the first steps in such a broad, long-term interdisciplinary effort that will help bridge the communication gap between astronomy and other disciplines. These concepts, and the approaches derived from them, will help to provide practical ways of analysis and visualization of the increasingly large and complex data sets. Such sophisticated new methods will also help to pave the way for effective automated analysis and processing of giant, complex data sets.

Acknowledgments

The authors would like to thank Alanna Connors for stimulating discussions and useful suggestions. They would also like to thank the referee for constructive suggestions that led to substantial improvements of the article. The first author would like to thank Michael Werner for support. This work was carried out with funding from the National Geospatial-Intelligence Agency University Research Initiative (NURI), Grant HM1582-08-1-0019, and support from NASA to the California Institute of Technology and the Jet Propulsion Laboratory.

References

- [1] D. York, J. Adelman, J. Anderson, et al., "The Sloan digital sky survey: technical summary," *Astronomical Journal*, vol. 120, pp. 1579–1587, 2000.
- [2] J. Tyson, C. Roat, J. Bosch, and D. Wittman, "LSST and the dark sector: image processing challenges," in *Proceedings of the 17th Annual Conference on Astronomical Data Analysis Software and Systems (ADASS '08)*, J. Lewis, R. Argyle, P. Bunclark, et al., Eds., vol. 394 of *ASP Conference Series*, ASP, London, UK, 2008.
- [3] A. S. Szalay, J. Gray, and J. Van den Berg, "Petabyte scale data mining: dream or reality?" in *Survey and Other Telescope Technologies and Discoveries*, vol. 4836 of *Proceedings of SPIE*, pp. 333–338, 2002.
- [4] M. A. C. Perryman, "GAIA: an astrometric and photometric survey of our galaxy," *Astrophysics and Space Science*, vol. 280, no. 1–2, pp. 1–10, 2002.
- [5] K. W. Hodapp, N. Kaiser, H. Aussel, et al., "Design of the Pan-STARRS telescopes," *Astronomische Nachrichten*, vol. 325, no. 6–8, pp. 636–642, 2004.
- [6] P. Price, E. Magnier, and J. Hobbitt, "Processing petabytes for Pan-STARRS, AAS Meeting 210," *Bulletin of the American Astronomical Society*, vol. 38, p. 183, 2007.
- [7] N. M. Law, S. R. Kulkarni, R. G. Dekany, et al., "The palomar transient factory: system overview, performance, and first results," *Publications of the Astronomical Society of the Pacific*, vol. 121, no. 886, pp. 1395–1408, 2009.
- [8] G. Zhao, Y. Q. Chen, J.-R. Shi, et al., "Stellar abundance and Galactic chemical evolution through LAMOST spectroscopic survey," *Chinese Journal of Astronomy and Astrophysics*, vol. 6, no. 3, pp. 265–280, 2006.
- [9] S. G. Djorgovski, C. Baltay, A. A. Mahabal, et al., "The Palomar-Quest digital synoptic sky survey," *Astronomische Nachrichten*, vol. 329, no. 3, pp. 263–265, 2008.
- [10] S. C. Barden and T. Armandroff, "Performance of the WIYN fiber-fed MOS system: Hydra," in *Fiber Optics in Astronomical Applications*, S. Barden, Ed., vol. 2476 of *Proceedings of SPIE*, pp. 56–67, Orlando, Fla, USA, April 1995.
- [11] B. C. Bigelow, A. M. Dressler, S. A. Shectman, and H. W. Epps, "IMACS: the multi-object spectrograph and imager for the Magellan I telescope," in *Optical Astronomical Instrumentation*, S. D'Odorico, Ed., vol. 3355 of *Proceedings of SPIE*, pp. 225–231, 1998.
- [12] I. J. Lewis, R. D. Cannon, K. Taylor, et al., "The Anglo-Australian observatory 2dF facility," *Monthly Notices of the Royal Astronomical Society*, vol. 333, no. 2, pp. 279–298, 2002.
- [13] R. M. Sharples, R. Bender, M. Lehnert, et al., "KMOS: an infrared multiple object integral field spectrograph for the ESO VLT," in *Ground-Based Instrumentation for Astronomy*, A. F. M. Moorwood and I. Masanori, Eds., vol. 5492 of *Proceedings of SPIE*, pp. 1179–1186, 2004.
- [14] R. Mark Wagner, "An overview of instrumentation for the Large Binocular Telescope," in *Ground-Based and Airborne Instrumentation for Astronomy*, I. S. McLean and I. Masanori, Eds., vol. 6269 of *Proceedings of SPIE*, 2006.
- [15] K. Freeman and J. Bland-Hawthorne, "Panoramic high resolution spectroscopy," in *Proceedings of the Panoramic Views of Galaxy Formation and Evolution*, T. Kodama, T. Yamada, and K. Aoki, Eds., vol. 399 of *ASP Conference Series*, p. 439, 2008.
- [16] S. Eto, T. Maihara, K. Ohta, et al., "The fiber multi-object spectrograph (FMOS) for the Subaru Telescope III," in *Ground-Based Instrumentation for Astronomy*, A. F. M. Moorwood and I. Masanori, Eds., vol. 5492 of *Proceedings of SPIE*, pp. 1314–1318, 2004.
- [17] S. C. Barden, A. Dey, B. Boyle, and K. Glazebrook, "KAOS: kilo-aperture optical spectrograph," in *Ground-Based Instrumentation for Astronomy*, A. F. M. Moorwood and I. Masanori, Eds., vol. 5492 of *Proceedings of SPIE*, pp. 364–372, Glasgow, UK, June 2004.
- [18] T. D. Ditto and J. M. Ritter, "Million object spectrograph," in *Advanced Optical and Mechanical Technologies in Telescopes*

- and *Instrumentation*, E. Atad-Ettinger and D. Lemke, Eds., vol. 7018 of *Proceedings of SPIE*, pp. 70182R.1–70182R.12, 2008.
- [19] M. Zaldarriaga and U. Seljak, “All-sky analysis of polarization in the microwave background,” *Physical Review D*, vol. 55, no. 4, pp. 1830–1840, 1997.
 - [20] R. Frazin, M. Jacob, W. Manchester, H. Morgan, and M. Wakin, “Toward reconstruction of coronal mass ejection density from only three points of view,” *Astrophysical Journal*, vol. 695, no. 1, pp. 636–641, 2009.
 - [21] M. Tegmark and M. Zaldarriaga, “Fast Fourier transform telescope,” *Physical Review D*, vol. 79, no. 8, Article ID 083530, 2009.
 - [22] T. Loredó, A. Accomazzi, J. Bloom, et al., “The astronomical information sciences: a keystone for 21st-century astronomy,” submitted to the *The Astronomy and Astrophysics Decadal Survey*.
 - [23] K. Borne, “Scientific data mining in astronomy,” in *Next Generation of Data Mining*, H. Kargupta, J. Han, P. S. Yu, R. Motwani, and V. Kumar, Eds., pp. 91–114, CRC Press, 2009.
 - [24] H. Ferguson, P. Greenfield, T. Axelrod, et al., “Astronomical data reduction and analysis for the next decade,” submitted to the *The Astronomy and Astrophysics Decadal Survey*.
 - [25] M. Grand, *Patterns in Java v.1*, Addison-Wesley, New York, NY, USA, 1998.
 - [26] G. Booch, J. Rumbaugh, and I. Jacobson, *The Unified Modeling Language User Guide*, Addison-Wesley, Reading, Mass, USA, 2005.
 - [27] M. Fayad, R. Johnson, and D. Schmidt, *Building Application Frameworks: Object Oriented Foundations of Framework Design*, John Wiley & Sons, New York, NY, USA, 1999.
 - [28] M. Pesenson, W. Roby, and B. McCollum, “Multiscale astronomical image processing based on nonlinear partial differential equations,” *Astrophysical Journal*, vol. 683, no. 1, pp. 566–576, 2008.
 - [29] K. France, S. McCandliss, and R. Lupu, “A Cometary Bow Shock and Mid-Infrared Emission Variations Revealed in *Spitzer* Observations of HD 34078 and IC 405,” *The Astrophysical Journal*, vol. 655, no. 2, pp. 920–939, 2007.
 - [30] A. Bovik, Ed., *Handbook of Image and Video Processing*, Elsevier, Amsterdam, The Netherlands, 2005.
 - [31] N. Paragios, Y. Chen, and O. Faugeras, *Handbook of Mathematical Models in Computer Vision*, Springer, Berlin, Germany, 2006.
 - [32] G. Herbig, “The spectra of Be- and Ae-TYPE stars associated with nebulosity,” *Astrophysical Journal Supplement*, vol. 4, p. 337, 1960.
 - [33] E. D. Feigelson and T. Montmerle, “High-energy processes in young stellar objects,” *Annual Review of Astronomy and Astrophysics*, vol. 37, no. 1, pp. 363–408, 1999.
 - [34] D. H. Clark and P. Murdin, “An unusual emission-line star/X-ray source/radio star, possibly associated with an SNR,” *Nature*, vol. 276, no. 5683, pp. 44–45, 1978.
 - [35] P. Feldman, C. Purton, T. Stiff, and S. Kwok, “H-alpha-Emission Object SS 433,” *IAU Circulars*, no. 3258, 1978.
 - [36] B. Margon, H. Ford, J. Katz, et al., “The bizarre spectrum of SS 433,” *Astrophysical Journal*, vol. 230, pp. L41–L45, 1979.
 - [37] B. A. Whitney, M. Sewilo, R. Indebetouw, et al., “*Spitzer* sage survey of the large magellanic cloud. III. Star formation and 1000 new candidate young stellar objects,” *The Astronomical Journal*, vol. 136, no. 1, pp. 18–43, 2008.
 - [38] W. van der Veen, H. Habing, and T. Geballe, “Objects in transition from the AGB to the planetary Nebula stage,” *Astronomy & Astrophysics*, vol. 226, pp. 108–136, 1989.
 - [39] W. van der Veen and H. Habing, “Far evolved AGB stars in the galactic bulge,” *Astronomy & Astrophysics*, vol. 231, no. 2, pp. 404–418, 1990.
 - [40] K. Finlator, Z. Ivezić, X. Fan, et al., “Optical and infrared colors of stars observed by the Two Micron All Sky Survey and the Sloan Digital Sky Survey,” *The Astronomical Journal*, vol. 120, no. 5, pp. 2615–2626, 2000.
 - [41] M. P. Egan, S. D. Van Dyk, and S. D. Price, “MSX, 2Mass, and the large magellanic cloud: a combined near- and mid-infrared view,” *The Astronomical Journal*, vol. 122, no. 4, pp. 1844–1860, 2001.
 - [42] R. Ortiz, S. Lorenz-Martins, W. J. Maciel, and E. M. Rangel, “Evolution from AGB to planetary nebula in the MSX survey,” *Astronomy & Astrophysics*, vol. 431, no. 2, pp. 565–574, 2005.
 - [43] S. V. Ramirez, R. G. Arendt, K. Sellgren, et al., “Point sources from a *Spitzer* IRAC survey of the galactic center,” *Astrophysical Journal*, vol. 175, no. 1, pp. 147–164, 2008.
 - [44] U. Munari, A. Henden, S. Kiyota, et al., “The mysterious eruption of V838 Mon,” *Astronomy & Astrophysics*, vol. 389, no. 2, pp. L51–L56, 2002.
 - [45] H. E. Bond, A. Henden, Z. G. Levay, et al., “An energetic stellar outburst accompanied by circumstellar light echoes,” *Nature*, vol. 422, no. 6930, pp. 405–408, 2003.
 - [46] J. A. Combi, J. F. Albacete-Colombo, and J. Martí, “A new search strategy for microquasar candidates using NVSS/2MASS and XMM-Newton data,” *Astronomy & Astrophysics*, vol. 477, no. 1, pp. 125–132, 2008.
 - [47] P. Brosche, “The manifold of galaxies. Galaxies with known dynamical parameters,” *Astronomy & Astrophysics*, vol. 23, pp. 259–268, 1973.
 - [48] S. Djorgovski and M. Davis, “Fundamental properties of elliptical galaxies,” *Astrophysical Journal*, vol. 313, p. 59, 1987.
 - [49] S. Djorgovski, R. Brunner, A. Mahabal, R. Williams, R. Granat, and P. Stolorz, “Challenges for cluster analysis in a virtual observatory,” in *Statistical Challenges in Astronomy*, E. Feigelson and G. Babu, Eds., pp. 127–141, Springer, Berlin, Germany, 2003.
 - [50] S. Haykin, *Neural Networks and Learning Machines*, Prentice Hall, Upper Saddle River, NJ, USA, 2009.
 - [51] N. Ball and R. Brunner, “Data mining and machine learning in astronomy,” *International Journal of Modern Physics D*, vol. 18, no. 17, pp. 1–53, 2009.
 - [52] G. Longo, C. Donalek, and G. Raiconi, “Artificial intelligence tools for data mining in large astronomical databases,” in *Globular Clusters-Guides to Galaxies*, p. 1431, Springer, Berlin, Germany, 2004.
 - [53] R. Bellman, *Adaptive Control Processes*, Princeton University Press, Princeton, NJ, USA, 1961.
 - [54] D. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, NY, USA, 1992.
 - [55] J. Matousek, *Lectures on Discrete Geometry*, Springer, New York, NY, USA, 2002.
 - [56] A. Bijaoui, “Astronomical application of information compression,” *Astronomy & Astrophysics*, vol. 30, p. 199, 1974.
 - [57] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York, NY, USA, 2005.

- [58] T. A. Boroson and R. F. Green, "The emission-line properties of low-redshift quasi-stellar objects," *Astrophysical Journal*, vol. 80, no. 1, pp. 109–135, 1992.
- [59] P. J. Francis, P. C. Hewett, C. B. Foltz, and F. H. Chaffee, "An objective classification scheme for QSO spectra," *Astrophysical Journal*, vol. 398, no. 2, pp. 476–490, 1992.
- [60] P. Francis and B. Wills, "An introduction to PCA," in *Quasars and Cosmology*, G. Ferland and J. Baldwin, Eds., vol. 162 of *ASP Conference Series*, p. 363, ASP, San Francisco, Calif, USA, 1999.
- [61] A. J. Connolly, I. Csabai, A. S. Szalay, D. C. Koo, R. G. Kron, and J. A. Munn, "Slicing through multicolor space: galaxy redshifts from broadband photometry," *The Astronomical Journal*, vol. 110, no. 6, pp. 2655–2664, 1995.
- [62] K. Glazebrook, A. R. Offer, and K. Deeley, "Automatic redshift determination by use of principal component analysis. I. Fundamentals," *Astrophysical Journal*, vol. 492, no. 1, pp. 98–109, 1998.
- [63] C. W. Yip, A. J. Connolly, D. E. Vanden Berk, et al., "Spectral classification of quasars in the sloan digital sky survey: eigenspectra, redshift, and luminosity effects," *The Astronomical Journal*, vol. 128, no. 6, pp. 2603–2630, 2004.
- [64] V. Wild and P. C. Hewett, "Peering through the OH forest: a new technique to remove residual sky features from sloan digital sky survey spectra," *Monthly Notices of the Royal Astronomical Society*, vol. 358, no. 3, pp. 1083–1099, 2005.
- [65] V. Wild, G. Kauffmann, T. Heckman, et al., "Bursty stellar populations and obscured active galactic nuclei in galaxy bulges," *Monthly Notices of the Royal Astronomical Society*, vol. 381, no. 2, pp. 543–572, 2007.
- [66] J. Kuraszkiewicz, B. Wilkes, G. Schmidt, and P. Smith, "Principal component analysis of the spectral energy distribution and emission line properties of red 2Mass active galactic nuclei," *Astrophysical Journal*, vol. 692, no. 2, pp. 1180–1189, 2009.
- [67] J. Gross and J. Yellen, Eds., *Handbook of Graph Theory*, CRC Press, Boca Raton, Fla, USA, 2004.
- [68] F. Chung, "Spectral graph theory," in *Conference Board of the Mathematical Sciences*, American Mathematical Society, Providence, RI, USA, 1997.
- [69] I. Pesenson, "Poincaré-type inequalities and reconstruction of Paley-Wiener functions on manifolds," *Journal of Geometric Analysis*, vol. 14, no. 1, pp. 101–121, 2004.
- [70] I. Pesenson, "An approach to spectral problems on Riemannian manifolds," *Pacific Journal of Mathematics*, vol. 215, no. 1, pp. 183–199, 2004.
- [71] I. Pesenson, "Variational splines on Riemannian manifolds with applications to integral geometry," *Advances in Applied Mathematics*, vol. 33, no. 3, pp. 548–572, 2004.
- [72] I. Pesenson, "Band limited functions on quantum graphs," *Proceedings of the American Mathematical Society*, vol. 133, no. 12, pp. 3647–3655, 2005.
- [73] I. Pesenson, "Polynomial splines and eigenvalue approximations on quantum graphs," *Journal of Approximation Theory*, vol. 135, no. 2, pp. 203–220, 2005.
- [74] I. Pesenson, "Analysis of band-limited functions on quantum graphs," *Applied and Computational Harmonic Analysis*, vol. 21, no. 2, pp. 230–244, 2006.
- [75] I. Pesenson, "Frames for spaces of Paley-Wiener functions on Riemannian manifolds, integral geometry and tomography," in *Contemporary Mathematics*, vol. 405, pp. 135–148, American Mathematical Society, Providence, RI, USA, 2006.
- [76] I. Pesenson, "Sampling in Paley-Wiener spaces on combinatorial graphs," *Transactions of the American Mathematical Society*, vol. 360, no. 10, pp. 5603–5627, 2008.
- [77] I. Pesenson, "Variational splines and Paley-Wiener spaces on combinatorial graphs," *Constructive Approximation*, vol. 29, no. 1, pp. 1–21, 2009.
- [78] I. Pesenson, "Paley-Wiener approximations and multiscale approximations in Sobolev and Besov spaces on manifolds," *Journal of Geometric Analysis*, vol. 19, no. 2, pp. 390–419, 2009.
- [79] I. Pesenson and M. Pesenson, "Sampling of bandlimited functions on combinatorial graphs," in *Proceedings of the 8th International Conference on Sampling Theory and Applications*, Marseille, France, 2009.
- [80] M. Pesenson, I. Pesenson, and B. McCollum, "Facing 'the Curse of Dimensionality': image fusion and nonlinear dimensionality reduction for advanced data mining and visualization of astronomical images," in *Proceedings of the 214th AAS Meeting*, 2009.
- [81] I. Pesenson and M. Pesenson, "Eigenmaps, minimal and bandlimited immersions of graphs into Euclidean spaces," *Journal of Mathematical Analysis and Applications*, vol. 366, no. 1, pp. 137–152, 2010.
- [82] D. Zhou, J. Huang, and B. Scholkopf, "Beyond pairwise classification and clustering using hypergraphs," Tech. Rep. 143, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2005.
- [83] E. Ihler, D. Wagner, and F. Wagner, "Modeling hypergraphs by graphs with the same min cut properties," *Information Processing Letters*, vol. 45, no. 4, pp. 171–175, 1993.
- [84] B. Mabelbrot, *Fractal Geometry of Nature*, W.H. Freeman, New York, NY, USA, 1977.
- [85] M. Schroeder, *Fractals, Chaos, Power Laws*, W. H. Freeman, New York, NY, USA, 1991.
- [86] C. Faloutsos, "Next generation data mining tools: power laws and self-similarity for graphs, streams and traditional data," in *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, vol. 2838 of *Lecture Notes in Artificial Intelligence*, pp. 10–15, 2003.
- [87] B. G. Elmegreen and D. M. Elmegreen, "Fractal structure in galactic star fields," *The Astronomical Journal*, vol. 121, no. 3, pp. 1507–1511, 2001.
- [88] R. Larson, "Towards understanding the stellar initial mass function," *Monthly Notices of the Royal Astronomical Society*, vol. 256, no. 4, pp. 641–646, 1992.
- [89] J. Einasto, M. Einasto, M. Gramann, and E. Saar, "Structure and formation of superclusters. XIII—the void probability function," *Monthly Notices of the Royal Astronomical Society*, vol. 248, p. 593, 1991.
- [90] T. Kurokawa, M. Morikawa, and H. Mouri, "Scaling analysis of galaxy distribution in the Las Campanas Redshift Survey data," *Astronomy & Astrophysics*, vol. 370, no. 2, pp. 358–364, 2001.
- [91] M. Bottorff and G. Ferland, "Fractal quasar clouds," *Astrophysical Journal*, vol. 549, no. 1, pp. 118–132, 2001.
- [92] E. M. de Gouveia dal Pino, M. Elisabete, A. Hetem, et al., "Evidence for a very large scale fractal structure in the universe from COBE measurements," *Astrophysical Journal*, vol. 442, no. 2, pp. L45–L48, 1995.
- [93] C. R. Ghezzi, E. M. de Gouveia dal Pino, and J. E. Horvath, "Asymmetric explosions of thermonuclear supernovae," *Monthly Notices of the Royal Astronomical Society*, vol. 348, no. 2, pp. 451–457, 2004.

- [94] B. G. Elmegreen and E. Falgarone, "A fractal origin for the mass spectrum of interstellar clouds," *Astrophysical Journal*, vol. 471, no. 2, pp. 816–821, 1996.
- [95] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, p. 7a, 2002.
- [96] T. Budavári, V. Wild, A. S. Szalay, L. Dobos, and C.-W. Yip, "Reliable eigenspectra for new generation surveys," *Monthly Notices of the Royal Astronomical Society*, vol. 394, no. 3, pp. 1496–1502, 2009.
- [97] D. Horn and A. Gottlieb, "Algorithm for data clustering in pattern recognition problems based on quantum mechanics," *Physical Review Letters*, vol. 88, no. 1, Article ID 018702, 4 pages, 2002.
- [98] M. Pesenson and I. Pesenson, "Image segmentation, unsupervised manifold learning and information visualization: a unified approach based on nonlinear dynamics," submitted.
- [99] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer, Berlin, Germany, 1984.
- [100] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization*, Cambridge University Press, Cambridge, UK, 2001.
- [101] C. Ware and R. Bobrow, "Supporting visual queries on medium sized node-link diagrams," *ACM Transactions on Applied Perception*, vol. 1, pp. 3–18, 2004.
- [102] E. Bertin and S. Arnouts, "SExtractor: software for source extraction," *Astronomy & Astrophysics*, vol. 117, no. 2, pp. 393–404, 1996.
- [103] J. Goodman and J. O'Rourke, Eds., *Handbook of Discrete and Computational Geometry*, Discrete Mathematics and Its Applications, Chapman & Hall/CRC, 2004.
- [104] E. Feigelson and J. Babu, *Statistical Challenges in Astronomy*, Springer, Berlin, Germany, 2003.
- [105] J. Silk, "An astronomer's perspective on SCMA III," in *Statistical Challenges in Astronomy, III*, E. D. Feigelson and G. J. Babu, Eds., p. 387, Springer, New York, NY, USA, 2003.
- [106] M. Gorski, E. Hivon, A. J. Banday, et al., "HEALPix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere," *Astrophysical Journal*, vol. 622, no. 2, pp. 759–771, 2005.
- [107] D. Pietrobon, A. Balbi, and D. Marinucci, "Integrated Sachs-Wolfe effect from the cross correlation of WMAP 3 year and the NRAO VLA sky survey data: New results and constraints on dark energy," *Physical Review D - Particles, Fields, Gravitation and Cosmology*, vol. 74, no. 4, Article ID 043524, 2006.
- [108] P. Baldi, G. Kerkycharian, D. Marinucci, and D. Picard, "High frequency asymptotics for wavelet-based tests for Gaussianity and isotropy on the torus," *Journal of Multivariate Analysis*, vol. 99, no. 4, pp. 606–636, 2008.
- [109] D. Marinucci, D. Pietrobon, A. Balbi, et al., "Spherical needlets for cosmic microwave background data analysis," *Monthly Notices of the Royal Astronomical Society*, vol. 383, no. 2, pp. 539–545, 2008.
- [110] G. Faÿ, F. Guillaoux, M. Betoule, J.-F. Cardoso, J. Delabrouille, and M. Le Jeune, "CMB power spectrum estimation using wavelets," *Physical Review D*, vol. 78, no. 8, Article ID 083013, 2008.
- [111] D. Pietrobon, A. Amblard, A. Balbi, P. Cabella, A. Cooray, and D. Marinucci, "Needlet detection of features in the WMAP CMB sky and the impact on anisotropies and hemispherical asymmetries," *Physical Review D*, vol. 78, no. 10, Article ID 103504, 2008.
- [112] D. Geller and D. Marinucci, "Spin wavelets on the sphere," <http://arxiv.org/abs/0811.2935>.
- [113] D. Geller, F. K. Hansen, D. Marinucci, G. Kerkycharian, and D. Picard, "Spin needlets for cosmic microwave background polarization data analysis," *Physical Review D*, vol. 78, no. 12, 2008.
- [114] D. Geller and A. Mayeli, "Continuous wavelets on compact manifolds," *Mathematische Zeitschrift*, vol. 262, no. 4, pp. 895–927, 2009.
- [115] I. Pesenson, "Lagrangian splines, spectral entire functions and Shannon-Whittaker theorem on manifolds," Temple University Research Report 95/87, 1995.
- [116] I. Pesenson, "A sampling theorem on homogeneous manifolds," *Transactions of the American Mathematical Society*, vol. 352, no. 9, pp. 4257–4269, 2000.
- [117] D. Geller and I. Pesenson, "Band-limited localized Parseval frames and Besov spaces on compact homogeneous manifolds," to appear in *Journal of Geometric Analysis*.
- [118] J.-L. Starck, F. Murtagh, and A. Bijaoui, *Image Processing and Data Analysis: The Multiscale Approach*, Cambridge University Press, Cambridge, UK, 1998.
- [119] J.-L. Starck and F. Murtagh, *Astronomical Image and Data Analysis*, Springer, Berlin, Germany, 2002.
- [120] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," in *Learning Theory*, vol. 3559 of *Lecture Notes in Computer Science*, pp. 486–500, Springer, Berlin, Germany, 2005.
- [121] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.
- [122] S. Lafon and A. B. Lee, "Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1393–1403, 2006.
- [123] J. Richards, P. Freeman, A. Lee, and C. Schafer, "Exploiting low-dimensional structure in astronomical spectra," *Astrophysical Journal*, vol. 691, pp. 32–42, 2009.
- [124] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [125] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [126] R. D'Abrusco, A. Staiano, G. Longo, et al., "Mining the SDSS archive. I. Photometric redshifts in the nearby universe," *Astrophysical Journal*, vol. 663, no. 2, pp. 752–764, 2007.
- [127] R. D'Abrusco, G. Longo, and N. A. Walton, "Quasar candidates selection in the Virtual Observatory era," *Monthly Notices of the Royal Astronomical Society*, vol. 396, no. 1, pp. 223–262, 2009.
- [128] F. Pasian, U. Becciani, and S. Cassisi, "VObs.it—the Italian Virtual Observatory," *Highlights of Astronomy*, vol. 14, p. 631, 2007.
- [129] M. Comparato, U. Becciani, A. Costa, et al., "Visualization, exploration, and data analysis of complex astrophysical data," *Publications of the Astronomical Society of the Pacific*, vol. 119, no. 858, pp. 898–913, 2007.
- [130] P. Draper, M. Currie, T. Jenness, et al., "GAIA 3D; visualizing data cubes," in *Proceedings of the 16th Annual Conference on Astronomical Data Analysis Software and Systems (ADASS '07)*, R. Shaw, F. Hill, and D. Bell, Eds., vol. 376 of *ASP Conference Series*, pp. 695–698, Tucson, Ariz, USA, 2007.

- [131] M. Borkin, A. Goodman, and M. Halle, "Application of medical imaging software to 3D visualization of astronomical data," in *Proceedings of the 16th Annual Conference on Astronomical Data Analysis Software and Systems (ADASS '07)*, R. Shaw, F. Hill, and D. Bell, Eds., vol. 376 of *ASP Conference Series*, pp. 621–624, Tucson, Ariz, USA, 2007.
- [132] W. L. Chan, K. Charan, D. Takhar, K. F. Kelly, R. G. Baraniuk, and D. M. Mittleman, "A single-pixel terahertz imaging system based on compressed sensing," *Applied Physics Letters*, vol. 93, no. 12, Article ID 121105, 2008.
- [133] J. Bobin, J.-L. Starck, and R. Ottensamer, "Compressed sensing in astronomy," *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 5, pp. 718–726, 2008.
- [134] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.
- [135] J. Romberg, "Imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 14–20, 2008.
- [136] Y. Wiaux, L. Jacques, G. Puy, A. M. M. Scaife, and P. Vanderghenst, "Compressed sensing imaging techniques for radio interferometry," *Monthly Notices of the Royal Astronomical Society*, vol. 395, no. 3, pp. 1733–1742, 2009.